Занятие 10. Регрессия (продолжение)

Вспоминаем предыдущее занятие

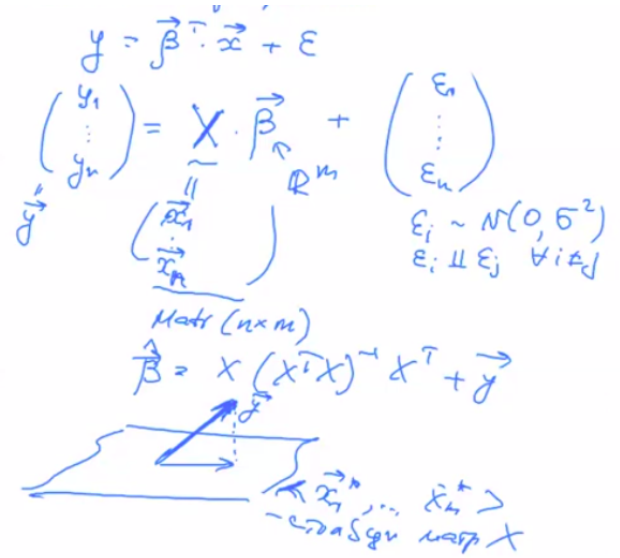
Данные:

$$(\bar{x}_1, y_1), \dots, (\bar{x}_n, y_n)$$

Нужно уметь предсказывать y по x.

Есть разные подходы:

- параметрические
  - линейная регрессия: $y = \bar{y}^\top \bar{x} + \varepsilon$



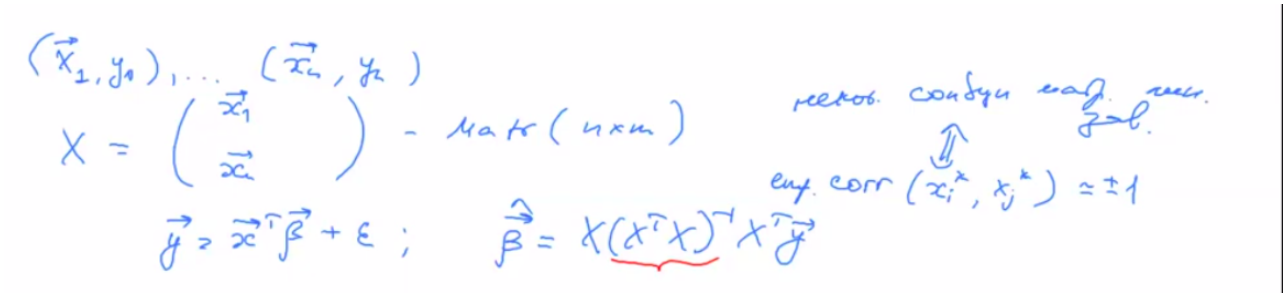
  - обобщенные линейные модели: есть экспоненциальное семейство $p(x, v)$. Предсказываем как $y_i \sim p(x, \bar{\beta}^\top \bar{x})$.
- непараметрические
  - Nadaraya-Watson (kernel regression)
  - регресионная (histogramm)
  - super smooth
  - LOESS
  - wavelets

Сегодня обсудим последние 4 пункта

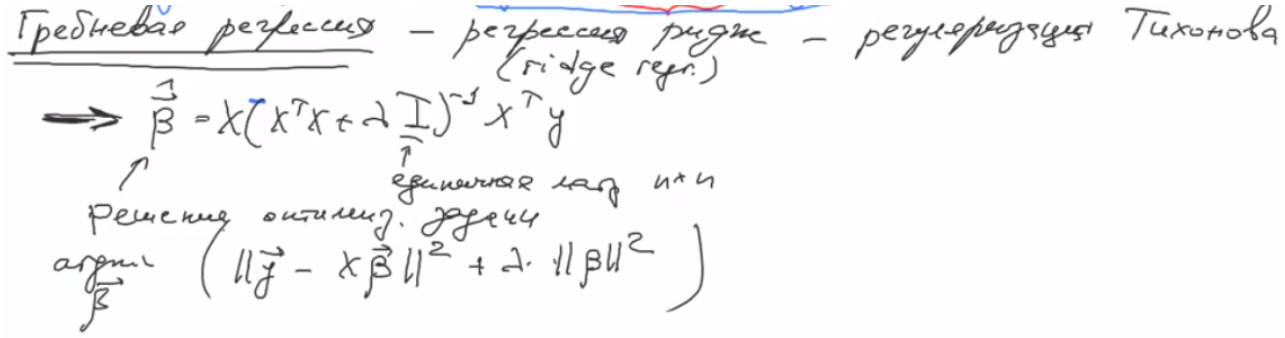
Пара слов про параметрическую регрессию

Гребневая регрессия

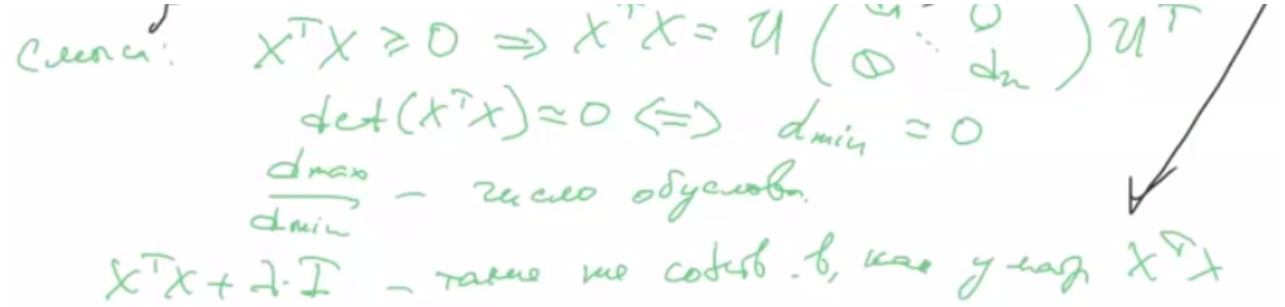
Если некоторые столбцы матрицы X линейно зависимы, то матрица $X^\top X$ будет необратима (или обратная к ней будет очень большой), а результат предсказания будет очень неустойчивым.



Решение:



Смысл: уменьшаем число обусловленности



То есть

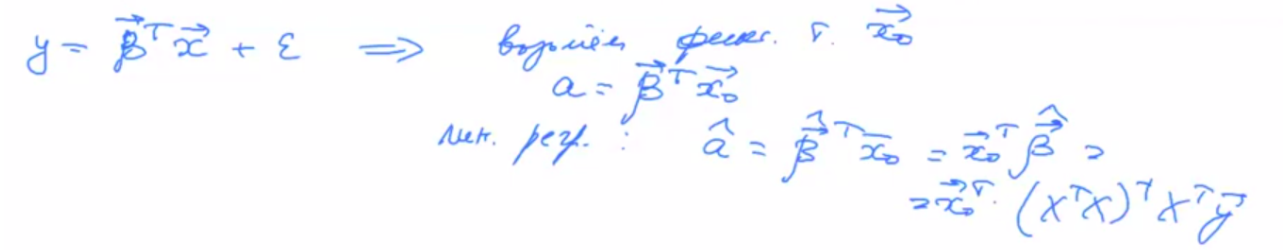
$$\mu(A) = \frac{\lambda_{\max}}{\lambda_{\min}} \mapsto \frac{\lambda_{\max} + \lambda}{\lambda_{\min} + \lambda} = \mu(A + \lambda I)$$

Это делает матрицу обратимой.

MSE в задаче регрессии

$$\mathrm{MSE}(\hat{a}) = \mathbb{E}[(\hat{a} - a)^2] = (\mathrm{bias}(\hat{a}))^2 + \mathrm{Var}\,\hat{a} = (\mathbb{E}\hat{a} - a)^2 + \mathrm{Var}\,\hat{a}$$

Для линейной регрессии имеем:



Теорема Гаусса-Маркова. Пусть $\mathbb{E}\varepsilon_i = 0,\ \ \mathrm{Var}\,\varepsilon_i = \sigma^2, \mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0$ (то есть допустимо, что $\varepsilon \nsim N(0, \sigma^2)$). Тогда

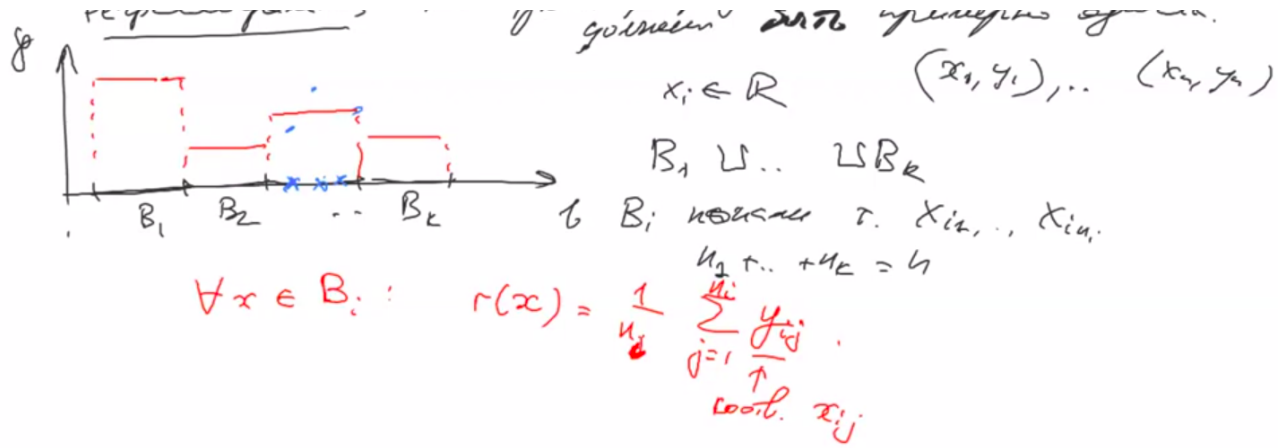
1. $\hat{a} = x_0^T (x^\top x)^{-1} x^\top y$ является несмещенной оценкой для $a = x_0^\top \beta$
2. Для любой другой несмещенной оценки дисперсия будет больше:

$$\forall \hat{a}^* : \mathbb{E}[\hat{a}^*] = a \quad \mathrm{Var}(\hat{a}^*) \geqslant \mathrm{Var}(\hat{a})$$

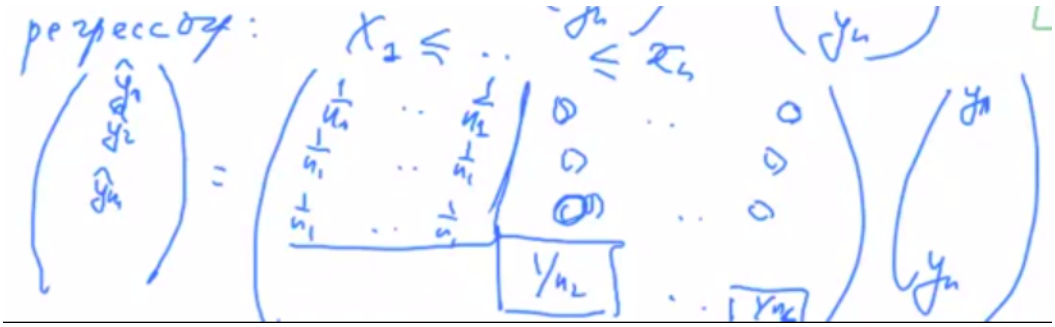
Непараметрическая регрессия

Регрессограмма

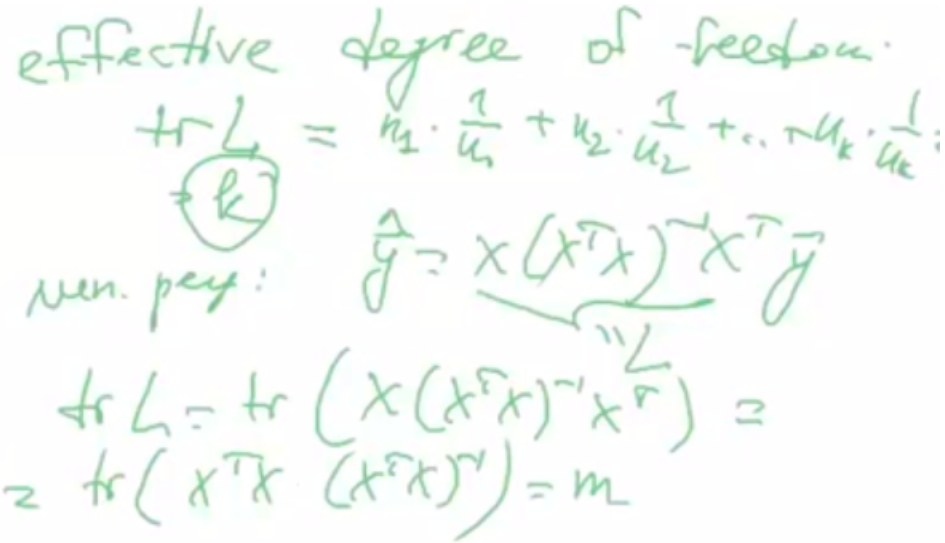
Значения регрессионной зависимости в близких точках должны быть примерно одинаковы



Linear smoother: $\hat{\bar{y}} = L \cdot \bar{y}$. В случае регрессограммы:



Effective degree of freedom:

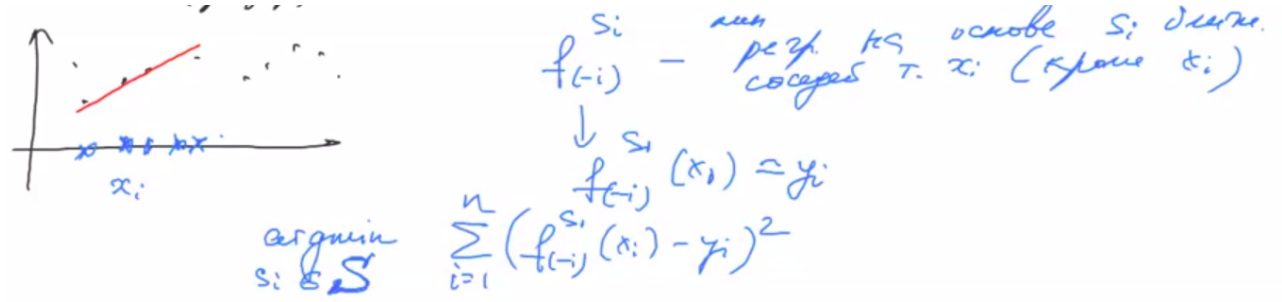


- в случае линейной регрессии равно количеству переменных
- в случае регрессограммы равно количеству блоков

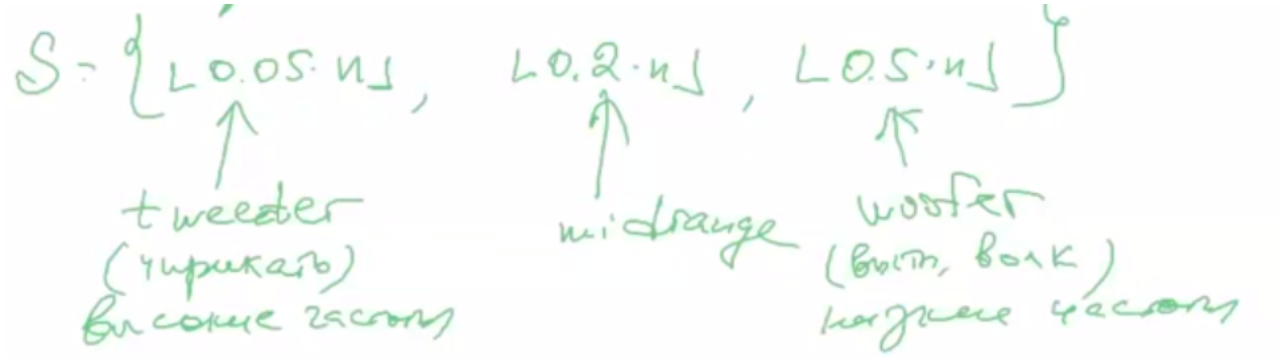
Supersmu

Super smoother – Friedman, 1984

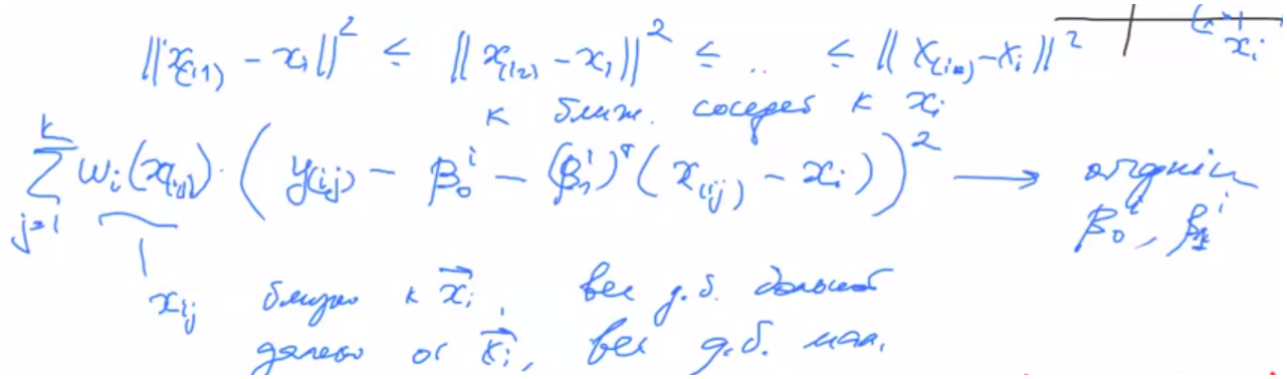
Философия: любая хорошая функция является локально линейной



Как выбирается S? Тут используется очень рандомный способ, но в целом рабочий. Множество S состоит из 3 чисел:

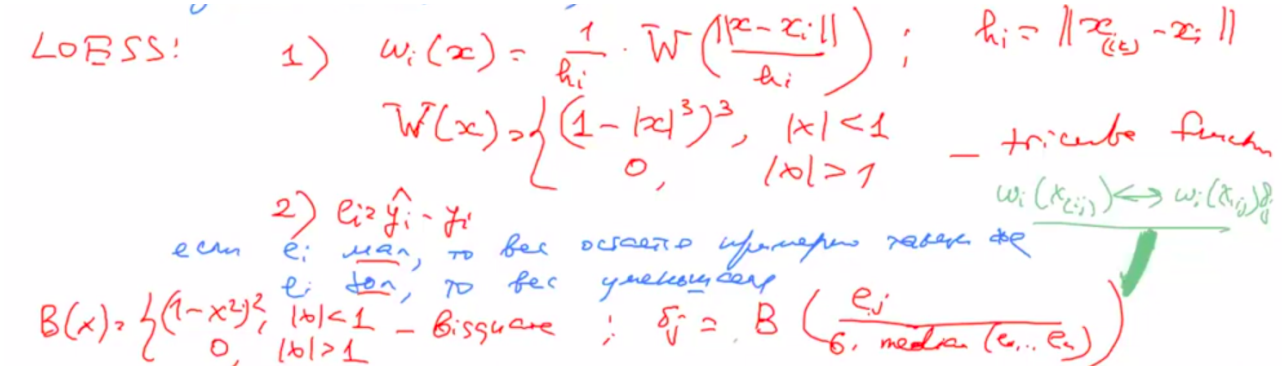


LOESS



Похоже на supersmu, но есть два отличия:

1. Добавляем каждому объекту вес
2. Пересчитываем вес на основании ошибки



Довольно похоже на AdaBoost

Wavelets

Напоминание:

$$L^2[a,b] = \{f : \int_a^b f^2(x)\,dx < \infty\}$$

В этом пространстве существует скалярное произведение:

$$\langle f, g\rangle = \int_a^b f(x)g(x)\,dx$$

Также тут есть бесконечномерный базис

$$\varphi_1, \dots, \varphi_n \in L^2[a,b]$$

для которого выполняется:

1. $\langle \varphi_i, \varphi_j\rangle = \delta_{ij}$ (ортонормированность)
2. $\forall i\ \langle f, \varphi_i\rangle = 0 \ \Rightarrow\ f \equiv 0$ (полная система)

Таким образом, любую функцию можно разложить по базису:

$$\forall f \in L^2[a,b]: \quad f(x) = \sum \alpha_i \varphi_i(x),\ \alpha_i = \langle f, \varphi_i\rangle$$

Известные базисы:

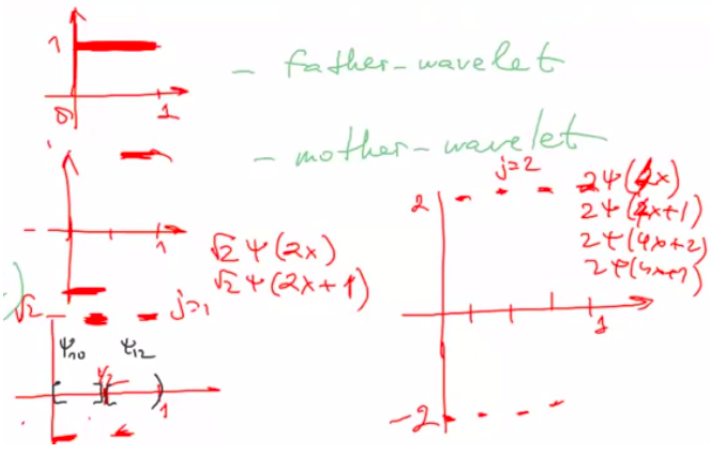
- полиномы Эрмита
- полиномы Лежандра
- базис Хаара (Haar basis)

Haar basis

Пусть $[a,b] = [0,1]$. Пусть

$$\varphi(x) \equiv 1 - \text{father wavelet}$$
$$\psi(x) = \begin{cases} -1, & x \in [0, 1/2] \\ 1, & x \in (1/2, 1] \end{cases} - \text{mother wavelet}$$
$$\psi_{jk}(x) = 2^{jk}\psi(2^j x + k), \quad j = 1, 2, \dots, \quad k = 0, 1, \dots 2^j - 1$$

Графики:



Разложим теперь функцию f по построенному базису:

$$f(x) = \alpha + \sum_{j=0}^{\infty}\sum_{k=0}^{2^j-1}\beta_{jk}\psi_{jk}(x), \quad \alpha = \int_0^1 f(x)\varphi(x)\,dx, \quad \beta_{jk} = \int_0^1 f(x)\psi_{jk}(x)\,dx$$

Применение к регрессии

$$y_i = f(x_i) + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \quad f \in L^2[0,1]$$

Оцениваем коэффициенты f в разложении по базису, предварительно обрубив ряд:

$$\hat{f}(x) = \hat{\alpha} + \sum_{j=-}^{J} \sum_{k=0}^{2^j-1} \hat{\beta}_{jk} \psi_{jk}(x), \quad \hat{\alpha} = \frac{1}{n} \sum_{i=1}^{n} y_i, \quad \hat{\beta}_{jk} = \frac{1}{n} \sum_{i=1}^{n} y_i \psi_{jk}(x_i)$$

При чем здесь нормальное распределение?

$$y_i \sim N(f(x_i), \sigma^2)$$
$$\Rightarrow \ \hat{\alpha} \sim N\Big(\underbrace{\frac{1}{n} \sum_{i=1}^{n} f(x_i)\varphi(x_i)}_{\to \int_0^1 f(x)\varphi(x)\,dx}, \underbrace{\frac{1}{n^2} \sum \varphi^2(x_i)\sigma^2}_{\approx \frac{1}{n}\int_0^1 \varphi^2(x)\,dx \sigma^2 \to 0} \Big)$$

Значит, при $n \to \infty$ наша оценка стремится к правильному значению

Сравнение регрессионных моделей

Generalized cross-validation (GCV)

Модель: $y_i = r(x_i) + \varepsilon_i$. Используем метод leave-one-out и строим $r_{(-i)}$. Можно перебрать все точки и получить общую ошибку $\sum (y_i - r_{(-i)}(x_i))^2$, но это долго. Вместо этого пользуются утверждением:

Утверждение.

$$y_i - r_{(-i)}(x) = \frac{y_i - r(x_i)}{1 - l_{ii}}$$

где l_{ii} это i-ый диагональный элемент матрицы $L = X(X^\top X)^{-1} X^\top$. Это нетривиальное утверждение (доказательство см. в блоге Roy Hyndman)

Тогда для *любого* линейного сглаживателя r (то есть $\hat{y} = Ly$) мы можем посчитать ошибку по формуле выше. В языке R используется несколько аппроксимаций:

$$\sum (y_i - r_{(-i)}(x_i))^2 = \sum_{i=1}^{n} \Big(\frac{y_i - r(x_i)}{1 - l_{ii}} \Big) \approx \Big\{ l_{11} + l_{nn} = \operatorname{tr} L \ \Rightarrow \ l_{ii} = \operatorname{tr} L / n \Big\} \approx \frac{\sum (y_i - r(x_i))}{(1 - \operatorname{tr} L/n)^2} \approx (1 + 2 \operatorname{tr} L/n) \sum (y_i - r(x_i))^2$$

Критерий Акаике

Тема из раздела model misspecification. У нас опять есть регрессионная модель: $y_i = r(x_i) + \varepsilon_i$. Будем предполагать, что $r(x_i) = \langle x_i, \beta \rangle$, $\varepsilon_i \sim N(0, \sigma^2)$. Считаем, что данные устроены как раз таким образом.

Теперь у нас есть модель-кандидат $\hat{r}$. С помощью KL-divergence можно измерить, насколько она близка к настоящей, посмотрев на распределение предсказаний:

$$KL(p_{true}, p_{cand}) = ?, \quad p_{true} = x\bar{\beta} + \bar{e} \sim N(x\bar{\beta}, \sigma^2 I), \quad p_{cand} \sim N(x\hat{\beta}, \hat{\sigma}^2 I)$$

Тогда

$$KL(p_{true}, p_{cand}) = n \log \hat{\sigma}^2 + \frac{n\sigma^2}{\hat{\sigma}^2} + \frac{1}{\sigma^2} (\mu - \hat{\mu})^\top (\mu - \hat{\mu}) + R(\mu, \sigma^2), \quad \mu = x\bar{\beta}$$

Однако мы не знаем μ, σ. Что делать? Возьмем матожидание для 2 и 3 слагаемого. Можно доказать, что

$$\frac{n\sigma^2}{\hat{\sigma}^2} \sim \mathcal{X}^2_{n-m}, \quad \frac{1}{\sigma^2} (\mu - \hat{\mu})^\top (\mu - \hat{\mu}) \sim F_{n, n-m}$$

Значит,

$$\mathbb{E}[\frac{n\sigma^2}{\hat{\sigma}^2}] = n - m, \quad \mathbb{E}[\frac{1}{\sigma^2} (\mu - \hat{\mu})^\top (\mu - \hat{\mu})] = \frac{nm}{n - m - 2}$$

В итоге получаем

$$L(p_{true}, p_{cand}) = n \log\left(\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2\right) + 1 + \frac{2(m+1)}{n+m-2} \to \min$$

Но что здесь m? Так же, как в случае GCV, заменим $m \mapsto \operatorname{tr} L$. Так можно сделать, ведь

$$\hat{y} = \underbrace{X(X^\top X)^{-1}X^\top}_{L} y, \quad \operatorname{tr} L = \operatorname{tr}(X^\top X)^{-1}(X^\top X) = \operatorname{tr} I = m$$

Анализ выживаемости

Теперь переходим к современным прикладным методам.

Анализ данных о моментах времени с некоторого определенного момента до наступления события. Особенности:

1. Есть цензурируемые наблюдения (отсутствует часть информации)
2. Несимметричные распределения ($\Rightarrow$ у них тяжелые хвосты)

Пример

Пациент сделал операцию от раке. После операции он ходит к доктору раз в несколько месяцев. Нам интересно, когда у него снова появятся симптомы болезни (рецидив). Основная проблема: пациент может перестать ходить к врачу (переезд в другой город / погиб в автокатосторфе / что-то еще). Из-за этого у нас неполные данные

Survival & hazard function

Survival funciton. Пусть T - время жизни пациента

$$s(T) = 1 - F_T(t) = 1 - \mathbb{P}\{T \leqslant t\} = \mathbb{P}\{T > t\}$$

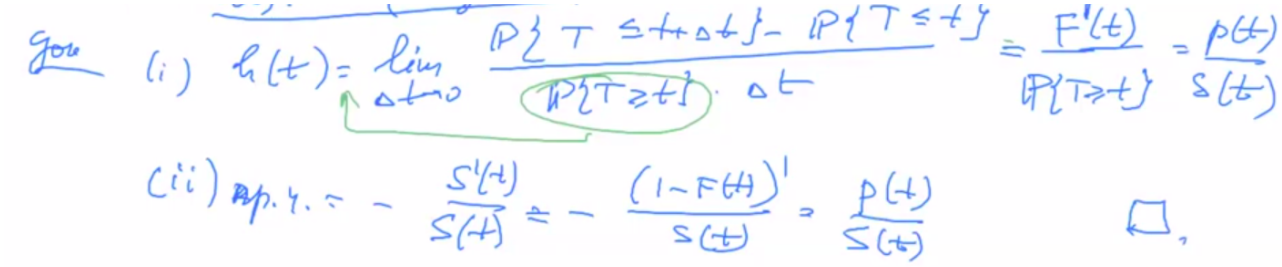
Hazard function (aka instance death rate):

$$h(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}\{t \leqslant T \leqslant t + \Delta t \mid T \geqslant t\}}{\Delta t}$$

Утверждение

1. $h(t) = p(t)/s(t)$
2. $h(t) = -(\log s(t))'$

Доказательство



Как оценить $h(t)$ и $s(t)$? Как использовать цензурируемые наблюдения?

Оценивание survival function

Банальный метод

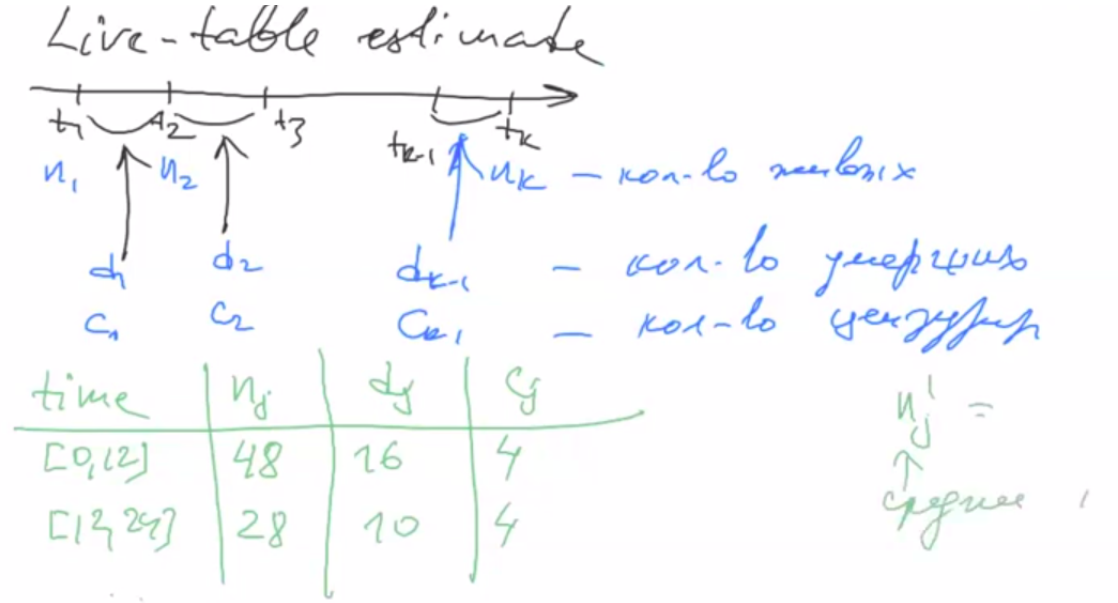
$$s(t) = 1 - F(t) \;\Rightarrow\; \hat{s}(t) = 1 - \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}[T_i \leqslant t] = \frac{\#\{T > t\}}{n}$$

Это плохой метод, т.к. мы не используем цензурируемые наблюдения

Хороший метод

Разобьем временную шкалу на бины и для каждого бина буем считать

- n_i - количество живых
- d_i - количество мертвых
- c_i - количество цензурируемых (которые последний раз встречались на этом интервале)



Пусть $n'_j = n_j - c_j/2$. Утверждается, что хорошей оценкой s будет

$$\hat{s}(t) = \prod_{j=1}^{k}(1 - \frac{d_j}{n'_j})$$

Почему?

$$s(t_k) = \mathbb{P}\{T > t_k\} = \mathbb{P}\{T > t_k \mid T > t_{k-1}\}\mathbb{P}\{T > t_{k-1}\} = \cdots = \prod_{j=1}^{k} \underbrace{\mathbb{P}\{T > t_j \mid T > t_{j-1}\}}_{\approx 1 - d_j/n'_j} \cdot \underbrace{\mathbb{P}\{T > t_0\}}_{=1}$$

У этого метода есть недостаток: при больших c_j у нас может возникнуть отрицательное число